ON SUBWORDS OF FORMAL LANGUAGES

by

G. Rozenberg

Institute of Applied Mathematics and Computer Science
University of Leiden
Wassenaarseweg 80, Leiden
The Netherlands

CU-CS-205-81 March 1981

ON SUBWORDS OF FORMAL LANGUAGES

G. Rozenberg
University of Leiden
The Netherlands

A way to understand the structure of a language is to investigate the set of all subwords that occur in the (words of the) language. A natural first step in such an investigation is simply to count the number of subwords of a given length in the language. Let for a language K, sub(K) denote the set of subwords of K, $sub_n(K)$ denote the number of subwords of length n occurring in K and let $\pi_K(n)$ denote the cardinality of $sub_n(K)$. Thus π_K is a function of positive integers assigning to each n the number of subwords of length n that occur in K; we refer to π_K as the subword complexity function of K. One can say that investigating the subword complexity of a language K forms a numerical approach to the investigation of the subwords of K.

In the first part of this paper we investigate the subword complexity of arbitrary languages. In particular we investigate to what extent a homomorphic mapping can influence the number of subwords.

Rather soon it becomes evident that to get a theory of subword complexity one has to consider languages that have "some structure" (as opposed to arbitrary languages). We choose to consider the class of languages generated by TOL systems and its subclasses. In the second part of this paper we demonstrate how the subword complexity (which is a global property in the sense that it is defined on a language independently of a system that generates it) of an TOL language is influenced by local restrictions (that is restrictions concerning the set of productions available) on an TOL system that generates it.

In the last part of this paper we consider global structural restrictions on the set of subwords of a given DOL language. For example we consider (following [8]) the restriction that no subword of a language is of the form xx where x is a non-empty word; such a language is called square-free. It turns out that the square-free condition on a DOL language restricts the number of possible subwords (of any length) quite considerably. In this way we see how a structural global restriction influences the global numerical measure.

This paper surveys results concerning subword complexity of formal languages obtained in the last few years. The proofs are not given, they can be found in the cited references.

PRELIMINARIES

We assume the reader to be familiar with the basic formal language theory. We use standard language theoretic notation and terminology. Perhaps the following points require an additional explanation. In this paper we consider finite alphabets only. On the other hand, since the problems considered become trivial otherwise, we consider infinite languages only (and consequently rewriting systems which generate infinite languages). For a finite set Z, #Z denotes its cardinality. For a word α, alph(α) denotes the set of all letters occurring in α and $|\alpha|$ denotes the length of α; Λ denotes the empty word. A word α is a subword of a word β if $\beta = \gamma\alpha\delta$ for some words γ,δ. For a language K, sub(K) denotes the set of all subwords (occurring in the words) of K and $\text{sub}_n(K)$ denotes the set of subwords of K of length n.

The following is the central notion of this paper. For a language K its subword complexity, denoted π_K, is the function of positive integers such that $\pi_K(n) = \#\text{sub}_n(K)$ for each positive integer n.

I. ARBITRARY LANGUAGES

In this section we investigate the subword complexity of arbitrary languages. First of all we establish the lower bound on the subword complexity of a language; we notice that there do not exist sublinear (but not constant) subword complexities.

Theorem 1. ([3]). Let K be a language. Either

(1). $\pi_K(n) \geq n+1$ for every positive integer n, or

(2). there exists a positive integer C such that $\pi_K(n) \leq C$ for every positive integer n. □

Then we turn to the investigation of the effect that a homomorphic mapping can have on a subword complexity. That is we investigate the relationship between $\pi_{h(K)}$ and π_K for a language K and a homomorphism h. It turns out that in general nothing meaningful can be said about this relationship.

Theorem 2. ([3]). For every positive integer e there exist alphabets Δ,Σ, a positive integer C, a language $K \subseteq \Delta^*$ and a homomorphism $h : \Delta^* \to \Sigma^*$ such that $\#\Sigma = e$ and $\pi_K(n) \leq Cn$, $\pi_{h(K)}(n) = e^n$ for every positive integer n. □

Even if we restrict ourselves to Λ-free homomorphisms the situation is "quite bad": no polynomial upper bound exists for the ratio $\dfrac{\pi_{h(K)}(n)}{\pi_K(n)}$.

Theorem 3. ([3]). There exists a language K and a Λ-free homomorphism h such that for no polynomial f, $\pi_{h(K)}(n) \leq f(n)\pi_K(n)$ for all positive integers n. □

To get a reasonable upper bound one has to put some structure on a language K. A natural first step in this direction is to require that π_K is a nondecreasing function.

Theorem 4. ([3]). Let $K \subseteq \Delta^*$ be a language such that π_K is a nondecreasing function and let h be a Λ-free homomorphism on Δ^*. Then there exists a positive integer constant C such that, for every positive integer n, $\pi_{h(K)}(n) \leq Cn\pi_K(n)$. □

We would like to remark here that the above result is not true for arbitrary (not necessarily Λ-free) homomorphisms.

II. LANGUAGES GENERATED BY GRAMMARS; THE EFFECT OF LOCAL RESTRICTIONS

In this section we investigate the subword complexity of languages generated by grammars; we have chosen to investigate languages generated by TOL systems (see, e.g., [7]).

A TOL system is a triple $G = (\Delta, H, \omega)$ where Σ is an alphabet, H is a nonempty finite set of finite substitutions (called tables) on Δ (into the subsets of Δ^*) and ω, the axiom, is an element of Σ^*. If for $h \in H$ and $a \in \Delta$, $\alpha \in h(a)$ then we say that $a \to \alpha$ is a production in G. The language of G, denoted L(G), is defined by $L(G) = \{x \in \Delta^* : x = \omega \text{ or } x \in h_1 \ldots h_m(\omega) \text{ where } k \geq 1 \text{ and } h_1, \ldots, h_m \in H\}$; L(G) is called a TOL language. We say that G is a deterministic TOL system, abbreviated DTOL system, if for every $h \in H$ and every $a \in \Delta$, $\#h(a) = 1$; accordingly L(G) is called a DTOL language.

Clearly, the set of all words over an alphabet Δ is a TOL language, so nothing specific can be said about the subword complexity of TOL languages in general. However, it turns out that the subword complexity (which is a global feature of a language) is sensitive to various local restrictions (that is restrictions on the sets of productions available in TOL systems). First of all it turns out that the (effect of the) deterministic restriction on TOL systems can be "detected by" looking at the subword complexity of generated languages.

Theorem 5. ([1]). Let Δ be a finite alphabet such that $\#\Delta = m \geq 2$. If K is a DTOL language, $K \subseteq \Delta^*$ then $\lim_{n \to \infty} \frac{\pi_K(n)}{m^n} = 0$. □

If $G = (\Delta, H, \omega)$ is a DTOL system such that $\#H = 1$ then we say that G is a DOL system (and L(G) is a DOL language). In this case if $H = \{h\}$ then we specify G in the form (Δ, h, ω).

Again, the restriction of DTOL systems to systems with one table only has an effect on the subword complexity of generated languages.

Theorem 6. ([2], [6]). Let K be a DOL language. There exists a positive integer constant C such that $\pi_K(n) \leq Cn^2$ for every positive integer n. □

The above result yields the best upper bound because there exist DOL languages with a subword complexity of order n^2 ([2], [6]).

A natural local restriction on a DOL system is a restriction on the length of (the right-hand side of) productions. A DOL system $G = (\Delta,h,\omega)$ is called growing, abbreviated as a GDOL system, if $\alpha = h(a)$ for $a \in \Delta$ implies that $|\alpha| \geq 2$; L(G) is referred to as a GDOL language. A DOL system $G = (\Delta,h,\omega)$ is called uniformly growing, abbreviated as a UGDOL system, if there exists a $t \geq 2$ such that if $\alpha = h(a)$ for $a \in \Delta$ implies that $|\alpha| = t$; L(G) is referred to as a UGDOL language.

Theorem 7. ([2], [6]). Let K be a DOL language.
(i). If K is a GDOL language then there exists a positive integer C such that $\pi_K(n) \leq Cn \log_2 n$ for every positive integer n.
(ii). If K is a UGDOL language then there exists a positive integer C such that $\pi_K(n) \leq Cn$ for every positive integer n. □

Also the above results ((i) and (ii)) yield the best upper bounds for the subword complexity of GDOL and UGDOL languages ([2], [6]).

As far as the effect of homomorphisms on the subword complexity is concerned we have the following results.

Theorem 8. ([3]). Let $K \subseteq \Delta^*$ be a DOL language and let h be a homomorphism of Δ^*. There exists a positive integer constant C such that $\pi_{h(K)}(n) \leq Cn^2$ for every positive integer n. □

Theorem 9. ([3]). There exists a UGDOL language $K \subseteq \Delta^*$, a positive real C and a homomorphism h of Δ^* such that $\pi_{h(K)}(n) \geq Cn^2$ for every positive integer n. □

The situation looks quite different if one considers Λ-free homomorphisms.

Theorem 10. ([3]). Let $K \subseteq \Delta^*$ be a DOL language and let h be a Λ-free homomorphism of Δ^*.
(i). If K is a GDOL language then there exists a positive integer C such that $\pi_{h(K)}(n) \leq Cn \log_2 n$ for every positive integer n.
(ii). If K is a UGDOL language then there exists a positive integer C such that $\pi_{h(K)}(n) \leq Cn$ for every positive integer n. □

III. LANGUAGES GENERATED BY GRAMMARS; THE EFFECT OF GLOBAL RESTRICTIONS

In this section we consider "structural" restrictions on the distribution of subwords in a DOL language.

Following [8] we say that a word is square-free if it does not have a subword of the form xx where x is a nonempty word. A language is called square-free if it

consists of square-free words only. We will consider square-free DOL languages now. Clearly, the square-free restriction is a global restriction (its formulation is independent on a DOL system that generates the DOL language under consideration). It is also a structural restriction in the sense that it talks about the structure of (the distribution of) subwords in words of a language. Again: also this restriction can be detected by the subword complexity function.

Theorem 11. ([4]). Let K be a square-free DOL language. There exists a positive integer C such that $\pi_K(n) \leq Cn \log_2 n$ for every positive integer n. □

Theorem 12. ([4]). There exist a square-free DOL language K and a positive integer constant D such that $\pi_K(n) \geq Dn \log_2 n$ for every positive integer n. □

To put the above results in a proper perspective we report also the following two results (the first of which is stated for arbitrary languages).

Theorem 13. ([4]). If K is a square-free language then $\pi_K(n) \geq n$ for every positive integer n. □

Theorem 14. ([4]). There exists a square-free DOL language K and a positive integer constant C such that $\pi_K(n) \leq Cn$ for every positive integer n. □

Another type of a global structural restriction is the following one. We say that a language K has a constant distribution if there exist an alphabet Δ and a positive integer constant C such that $\underline{alph}(\alpha) = \Delta$ for every word in $\underline{sub}_C(K)$. Also this structural global restriction is detectable by the subword complexity function.

Theorem 15. ([5]). Let K be a DOL language that has a constant distribution. There exists a positive integer constant C such that $\pi_K(n) \leq Cn$ for every positive integer n. □

ACKNOWLEDGEMENTS.

The author is deeply indebted to A. Ehrenfeucht for the continuous cooperation in the research reported. The author gratefully acknowledges the support by NSF grant MCS 79-03838

REFERENCES

1. A. Ehrenfeucht and G. Rozenberg, A limit for sets of subwords in deterministic TOL systems, Information Processing Letters, 2 (1973), 70-73.
2. A. Ehrenfeucht, K.P. Lee and G. Rozenberg, Subword complexities of various classes of deterministic developmental languages without interactions, Theoretical Computer Science, 1 (1975), 59-76.
3. A. Ehrenfeucht and G. Rozenberg, On subword complexities of homomorphic images of languages, Rev. Fr. Automat. Inform. Rech. Opér., Ser. Rouge, to appear.
4. A. Ehrenfeucht and G. Rozenberg, On the subword complexity of square-free DOL languages, Theoretical Computer Science, to appear.

5. A. Ehrenfeucht and G. Rozenberg, On the subword complexity of DOL languages with a constant distribution, Institute of Applied Mathematics and Computer Science, University of Leiden, The Netherlands, Technical Report No. 81-21, 1981.
6. K.P. Lee, Subwords of developmental languages, Ph.D. thesis, State University of New York at Buffalo, 1975.
7. G. Rozenberg and A. Salomaa, The mathematical theory of L Systems, Academic Press, London, New York, 1980.
8. A. Thue, Über unendliche Zeichenreihen, Norsk. Vid. Selsk. Skr. I Mat.-Nat.Kl. 7 (1906), 1-22.